Recent research from UChicago shows that Large Language Models (LLMs) excel at analyzing written documents such as conference call transcripts and disclosures, and using them to generate informative summaries and accurate predictions of risk and investment. In this paper, the authors test whether LLMs can also process quantitative material; specifically, they study GPT 4.0 Turbo’s ability to analyze financial statements.

The authors collect and anonymize financial statements from 15,401 corporations from 1968 to 2021. They design and implement two prompts to instruct GPT to analyze the financial statements. In the first prompt, the authors simply instruct GPT to use the statements to determine whether the company’s earnings will increase or decrease in the following year. In the second prompt, the authors break down the problem into steps that parallel those followed by human analysts and ask the model to analyze financial statements. The authors compare GPT’s predictions to those issued by actual analysts in the month following the release of the financial statements. They find the following:

- GPT outperforms financial analysts at predicting earnings changes, on average. While the authors’ “simple” prompt tends to underperform human analysts, the chain-of-thought-style prompt designed to emulate human reasoning is accurate 60% of the time, seven percentage points better than human analysts’ average accuracy.

- GPT exhibits a relative advantage over human analysts in situations when human analysts tend to struggle. The authors identify instances in their data when human analysts are expected to exhibit bias and disagreement,
and show that in these instances, GPT forecasts are more informative in predicting the direction of future earnings.

- The authors confirm that GPT’s predictive accuracy is not owed to the model’s ability to identify the specific company represented by the anonymized financial statements and “remember” its actual earnings.
- GPT’s predictions are on par with those of narrowly trained machine learning models.
- Trading strategies based on GPT’s predictions yield higher *alphas* and *Sharpe ratios* than based on other machine learning algorithms.

The upshot is that LLMs can analyze numbers in financial statements without any narrative context. These findings should be of interest to investors and regulators, as they indicate the potential for LLMs to democratize financial information processing. For example, this research shows that generative AI is not merely a tool that can assist investors, but it can also play a more active role in helping consumers make informed decisions. More broadly, this research implies that LLMs may take a central role in financial decision-making in the future.

---

**BFI Blackboard**

**Alpha:** An investment’s performance compared to a market index or benchmark, indicating the value added or lost by a portfolio manager’s decisions.

**Sharpe ratio:** The risk-adjusted return on an investment, calculated by comparing its excess return to its standard deviation, indicating how much extra return investors get for extra risk.

---

**READ THE WORKING PAPER**

NO. 2024-65 · MAY 2024

**Financial Statement Analysis with Large Language Models**

bfi.uchicago.edu/working-paper/2024-65

---

**ABOUT OUR SCHOLARS**

**Alex G. Kim**  
PhD Student, Chicago Booth

**Maximilian Muhn**  
Assistant Professor of Accounting, Chicago Booth

**Valeri Nikolaev**  
James H. Lorie Professor of Accounting and FMC Faculty Scholar, Chicago Booth